

What is claimed is:

- 1 1. A system for identifying language attributes through probabilistic
2 analysis, comprising:
 - 3 a storage storing a set of language classes, which each identify a language
4 and a character set encoding, and a plurality of training documents;
 - 5 an attribute modeler evaluating occurrences of one or more document
6 properties within each training document and, for each language class, calculating
7 a probability for the document properties set conditioned on the occurrence of the
8 language class; and
 - 9 a text modeler evaluating byte occurrences within each training document
10 and, for each language class, calculating a probability for the byte occurrences
11 conditioned on the occurrence of the language class.
- 1 2. A system according to Claim 1, further comprising:
 - 2 a training engine calculating an overall probability for each language class
3 by evaluating the probability for the document properties set and the probability
4 for the byte occurrences.
- 1 3. A system according to Claim 1, further comprising:
 - 2 an assignment module assigning the overall probability for each language
3 class in accordance with the formula:
$$4 \quad \arg \max_{cls} P(text | cls) \cdot P(props | cls) \cdot P(cls)$$
- 5 where cls is the language class, $text$ is the byte occurrences set, $props$ are the
6 document properties, and $P(text | cls)$ is the probability for the byte occurrences,
7 and $P(props | cls)$ is the probability for the document properties set.
- 1 4. A system according to Claim 1, wherein the document properties
2 comprise at least one of top level domain, HTTP content character set encoding
3 and language header parameters, and HTML content character set encoding and
4 language metatags.

1 5. A system according to Claim 4, further comprising:
2 an assignment module assigning the probability for the document
3 properties set in accordance with the formula:

4 $P(tld, enc | cls) \cdot P(cls)$

5 where *tld* is the top level domain, *enc* is the character set encoding and *cls* is the
6 language class.

1 6. A system according to Claim 1, further comprising:
2 a counting module counting byte co-occurrences within each training
3 document, and determining the probability for the byte occurrences based on the
4 byte co-occurrences.

1 7. A system according to Claim 6, wherein the byte co-occurrences
2 comprise a set of trigrams, further comprising:
3 a probability module calculating a probability of each trigram as the
4 number of occurrences of the trigram divided by the total number of trigram
5 occurrences in each of the training documents for each language class.

1 8. A system according to Claim 7, further comprising:
2 an assignment module assigning the probability for the byte occurrences
3 set in accordance with the formula:

4 $P(text | cls)$

5 where *text* is the set of trigrams and *cls* is the language class.

1 9. A system according to Claim 1, further comprising:
2 a training engine performing iterative training by providing the probability
3 for the document properties set and the probability for the byte occurrences set
4 respectively to the evaluation of byte occurrences and assignment of the set of
5 language classes.

1 10. A system according to Claim 1, further comprising:

2 a back off module evaluating less frequently occurring document
3 properties by calculating a probability for each less frequently occurring
4 document property conditioned on the occurrence of the language class.

1 11. A system according to Claim 1, further comprising:
2 a plurality of unlabeled documents; and
3 a classifier classifying one or more unlabeled documents by at least one
4 language class, comprising evaluating occurrences of one or more document
5 properties within the unlabeled document, evaluating byte occurrences within the
6 unlabeled document, and assigning a probability for the document properties set
7 and the byte occurrences for the unlabeled document conditioned on the
8 occurrence of the language class.

1 12. A system according to Claim 11, further comprising:
2 a class selector selecting the at least one language class having a
3 substantially highest probability.

1 13. A system according to Claim 11, further comprising:
2 a probability threshold; and
3 a pruner pruning at least one language class falling below the probability
4 threshold.

1 14. A system according to Claim 1, wherein each training document
2 comprises one of a Web page and a news message.

1 15. A method for identifying language attributes through probabilistic
2 analysis, comprising:
3 defining a set of language classes, which each identify a language and a
4 character set encoding, and a plurality of training documents;
5 evaluating occurrences of one or more document properties within each
6 training document and, for each language class, calculating a probability for the
7 document properties set conditioned on the occurrence of the language class; and

evaluating byte occurrences within each training document and, for each language class, calculating a probability for the byte occurrences conditioned on the occurrence of the language class.

1 16. A method according to Claim 15, further comprising:
2 calculating an overall probability for each language class by evaluating the
3 probability for the document properties set and the probability for the byte
4 occurrences.

1 17. A method according to Claim 15, further comprising:
2 assigning the overall probability for each language class in accordance
3 with the formula:

$$4 \quad \arg \max_{cls} P(text | cls) \cdot P(props | cls) \cdot P(cls)$$

5 where cls is the language class, $text$ is the byte occurrences set, $props$ are the
 6 document properties, and $P(text | cls)$ is the probability for the byte occurrences,
 7 and $P(props | cls)$ is the probability for the document properties set.

1 18. A method according to Claim 15, wherein the document properties
2 comprise at least one of top level domain, HTTP content character set encoding
3 and language header parameters, and HTML content character set encoding and
4 language metatags.

1 19. A method according to Claim 18, further comprising:
2 assigning the probability for the document properties set in accordance
3 with the formula:

$$4 \quad \quad \quad P(tld, enc \mid cls) \cdot P(cls)$$

5 where *tld* is the top level domain, *enc* is the character set encoding and *cls* is the
6 language class.

1 20. A method according to Claim 15, further comprising:
2 counting byte co-occurrences within each training document; and

3 determining the probability for the byte occurrences based on the byte co-
4 occurrences.

1 21. A method according to Claim 20, wherein the byte co-occurrences
2 comprise a set of trigrams, further comprising:

3 calculating a probability of each trigram as the number of occurrences of
4 the trigram divided by the total number of trigram occurrences in each of the
5 training documents for each language class.

1 22. A method according to Claim 21, further comprising:

2 assigning the probability for the byte occurrences set in accordance with
3 the formula:

4
$$P(text | cls)$$

5 where *text* is the set of trigrams and *cls* is the language class.

1 23. A method according to Claim 15, further comprising:

2 performing iterative training by providing the probability for the document
3 properties set and the probability for the byte occurrences set respectively to the
4 evaluation of byte occurrences and assignment of the set of language classes.

1 24. A method according to Claim 15, further comprising:

2 evaluating less frequently occurring document properties by calculating a
3 probability for each less frequently occurring document property conditioned on
4 the occurrence of the language class.

1 25. A method according to Claim 15, further comprising:

2 accessing a plurality of unlabeled documents; and
3 classifying one or more unlabeled documents by at least one language
4 class, comprising:

5 evaluating occurrences of one or more document properties within
6 the unlabeled document;

7 evaluating byte occurrences within the unlabeled document; and

8 assigning a probability for the document properties set and the byte
9 occurrences for the unlabeled document conditioned on the occurrence of the
10 language class.

1 26. A method according to Claim 25, further comprising:
2 selecting the at least one language class having a substantially highest
3 probability.

1 27. A method according to Claim 25, further comprising:
2 defining a probability threshold; and
3 pruning at least one language class falling below the probability threshold.

1 28. A method according to Claim 15, wherein each training document
2 comprises one of a Web page and a news message.

1 29. A computer-readable storage medium holding code for performing
2 the method according to Claim 15.

- 1 30. A system for identifying documents by language using
- 2 probabilistic analysis of language attributes, comprising:
 - 3 a set of language classes, each language class comprising a language name
 - 4 and a character set encoding name;
 - 5 a training corpora comprising a plurality of training documents;
 - 6 an attribute modeler training an attribute model by evaluating a top level
 - 7 domain and character set encoding associated with each training document and,
 - 8 for each language class, calculating a probability for each such top level domain
 - 9 and character set encoding conditioned on the occurrence of the each language
 - 10 class; and
 - 11 a text modeler training a text model by evaluating co-occurrences of a
 - 12 plurality of bytes within each training document and, for each language class,
 - 13 calculating a probability for the byte co-occurrences conditioned on the
 - 14 occurrence of the each language class.

1 31. A system according to Claim 30, further comprising:

2 a training engine calculating an overall probability for each language class
3 by evaluating the probability for the top level domain and character set encoding
4 based on the attribute model and the probability for the byte occurrences based on
5 the text model.

1 32. A system according to Claim 31, further comprising:
2 a classifier classifying one or more documents, comprising:
3 an attribute evaluator evaluating a top level domain and character
4 set encoding in each document and applying the attribute model to the evaluated
5 top level domain and character set encoding;
6 a text evaluator evaluating byte occurrences in each document and
7 applying the text model to the evaluated byte occurrences; and
8 an assignment module assigning at least one language class based
9 on the applications of the attribute model and the text model.

1 33. A method for identifying documents by language using
2 probabilistic analysis of language attributes, comprising:
3 defining a set of language classes, each language class comprising a
4 language name and a character set encoding name;
5 assembling a training corpora comprising a plurality of training
6 documents;
7 training an attribute model by evaluating a top level domain and character
8 set encoding associated with each training document and, for each language class,
9 calculating a probability for each such top level domain and character set
10 encoding conditioned on the occurrence of the each language class; and
11 training a text model by evaluating co-occurrences of a plurality of bytes
12 within each training document and, for each language class, calculating a
13 probability for the byte co-occurrences conditioned on the occurrence of the each
14 language class.

1 34. A method according to Claim 33, further comprising:

2 calculating an overall probability for each language class by evaluating the
3 probability for the top level domain and character set encoding based on the
4 attribute model and the probability for the byte occurrences based on the text
5 model.

1 35. A method according to Claim 34, further comprising:
2 classifying one or more documents, comprising:
3 evaluating a top level domain and character set encoding in each
4 document and applying the attribute model to the evaluated top level domain and
5 character set encoding;
6 evaluating byte occurrences in each document and applying the
7 text model to the evaluated byte occurrences; and
8 assigning at least one language class based on the applications of
9 the attribute model and the text model.

1 36. A computer-readable storage medium holding code for performing
2 the method according to Claim 30.

1 37. An apparatus for identifying documents by language using
2 probabilistic analysis of language attributes, comprising:
3 means for defining a set of language classes, each language class
4 comprising a language name and a character set encoding name;
5 means for training an attribute model by assigning at least one top level
6 domain and character set encoding pairing to at least one language class for each
7 of a plurality of training documents and calculating a probability for each such top
8 level domain and character set encoding pairing conditioned on the occurrence of
9 the assigned language class; and
10 means for training a text model by evaluating co-occurrences of a plurality
11 of bytes within each training document and, for each language class, calculating a
12 probability for the byte co-occurrences conditioned on the occurrence of the
13 language class based on the attribute model.